SIMILARITY STRUCTURE IN PERCEPTUAL AND PHYSICAL MEASURES FOR VISUAL CONSONANTS ACROSS TALKERS

Jintao Jiang[1], *Abeer Alwan*[1], *Lynne E. Bernstein*[2], *Edward T. Auer, Jr*[2]., *and Patricia A. Keating*[3]

[1]Department of Electrical Engineering, UCLA, Los Angeles, CA 90095
[2]Department of Communication Neuroscience, House Ear Institute, Los Angeles, CA 90057
[3]Department of Linguistics, UCLA, Los Angeles, CA 90095
{jjt, alwan}@icsl.ucla.edu, {lbernstein, eauer}@hei.org, keating@humnet.ucla.edu

ABSTRACT

This paper investigates the relationship between visual confusion matrices and physical (facial) measures. The similarity structure in perceptual and physical measures for visual consonants was examined across four talkers. Four talkers, spanning a wide range of rated visual intelligibility, were recorded producing 69 Consonant-Vowel (CV) syllables. Audio, video, and 3-D face motion were recorded. Each talker's CV productions were presented for identification in a visual-only condition to six viewers with average or better lipreading ability. The obtained visual confusion matrices demonstrated that phonemic equivalence classes were related to visual intelligibility and were talker and vowel context dependent. Physical measures accounted for about 63% of the variance of visual consonant perception, with C/u/ syllables yielding higher correlations than C/a/ and C/i/ syllables.

1. INTRODUCTION

Potential applications of visual speech are now widely acknowledged,. yet, there is meager fundamental knowledge about optical phonetics. Acoustic phonetics has benefited from numerous experiments involving careful measurements on natural speech and from perceptual experiments using acoustic signals. On the contrary, only a few papers have examined the relationship between visual phonetic perception and optical phonetics. Finding physical cues to visual speech perception can help understand visual speech perception and ultimately improve visual speech synthesis.

Montgomery and Jackson [1] examined the relationship between visual vowel perception and physical characteristics in an experiment with four female talkers, ten viewers, and fifteen vowels in /hVg/ nonsense words. Since vowel pronunciation is relatively steady and of long duration in this context, the authors used a set of static descriptors to define physical characteristics: lip height, lip width, lip aperture, acoustic duration, and visual duration. Their results indicated that the physical measures were moderately successful as predictors of vowel perception (approximately 50% of the variance was accounted for) but the predictions were talker dependent. In [2], Kricos and Lesner examined differences in visual intelligibility across talkers. In [3], Benoît et al. showed that speech identification accuracy increased as more of the face (lip, jaw, cheek) was made visible. The authors of [2, 3] did not, however, examine the corresponding optical signals. In addition, most physical measures reported in the literature thus far have focused on talkers' lips.

In a previous study [4], we examined the predictability of visual consonant perception from physical measures which included a variety of measures of visible articulatory movements. Talker differences were not examined, even though the four talkers have different visual intelligibility ratings. In this study, we continue to study visible articulatory movements to examine how the relationship between visual perception and physical measures varies across talkers.

2. METHOD

2.1. Data recording

2.1.1. Talkers

Four native American English talkers (two males, M1 and M2, and two females, F1 and F2) with different sentence intelligibility ratings were recorded [5]. Visual intelligibility was tested on 20 sentences and judged by five deaf adults. The mean sentence intelligibility ratings were 3.6, 8.6, 1.0, and 6.6 for talker M1, M2, F1, and F2, respectively. These mean sentence intelligibility ratings are on a scale of 1-10 where 1 is not intelligible and 10 is very intelligible.

2.1.2. Stimuli

The speech material consisted of two repetitions of 69 CV syllables where the vowel was one of /a, i, u/ and the consonant was one of the 23 American English consonants, /y, w, r, l, m, n, p, t, k, b, d, g, h, □, □, s, z, f, v, □, □, t□, d□. Three data streams were recorded simultaneously and synchronized [5] for each CV production: acoustic (DAT), optical (BETACAM video), and 3-D motion (Qualisys). The video of the CV tokens was edited to a second BETACAM tape for presentation to perceivers.

2.1.3. Recording channels

Face motion was captured by a Qualysis™ motion capture system, which used an infrared flash to reflect light from reflectors glued on the face. Fig. 1 shows the number and placement of 20 optical reflectors, which were placed on the nose bridge (one), eye brows (two), lip contour (eight), chin (three), and cheeks (six). Reflectors 1, 2, and 3 were used for head motion compensation and were not used in the analyses.

0-7803-7402-9/02/$17.00 ©2002 IEEE

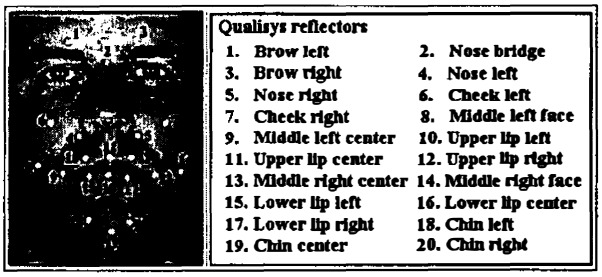

Figure 1. Placement of optical face reflectors.

2.2. Physical measures

2.2.1. Head motion compensation

Small head movements occurred during recordings. The reflectors on the nose bridge (2) and eye brows (1 and 3) were relatively stable across the session, and thus were used for head movement compensation by constructing a new 3-D coordinate system as shown in Fig. 2. Optical data were projected onto the new coordinate system, which represents a stable head structure.

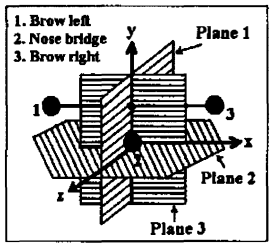

Figure 2. 3-D coordinate system for head motion compensation.

2.2.2. Physical distances between consonants

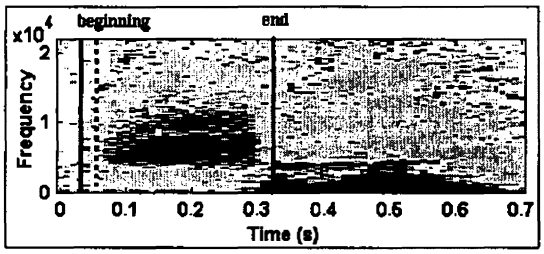

Figure 3. Consonant segment in a /Sa/ syllable.

The sampling frequency for the optical data was 120 Hz. Since we were interested in analyzing the consonants, only the initial part of the CV was used. Fig. 3 illustrates how this was done for /Sa/. Starting point detection was based on the acoustic signal because talkers sometimes produced unrelated movements well before the pronunciation, but the starting point of the audio signal was stable. Due to forward anticipation effects, that is face movements are ahead of the acoustic signals, a segment was defined 30 ms prior to the acoustic beginning of the CV (dashed line) and lasted for 280 ms (between the 2 solid lines). A segment had 34 optical frames (280 ms * 120 frame/s ≈ 34 frames). Each consonant segment should have included the preparation, dynamic movements, and CV transition. Each optical segment were then organized into matrices as follows:

$$O^{T,CV,\beta}_{(1:51,1:34)} = \begin{bmatrix} o_{1,1} & \cdots & o_{1,34} \\ \vdots & \vdots & \vdots \\ o_{51,1} & \cdots & o_{51,34} \end{bmatrix}, \quad (1)$$

where T, CV, and β, refer to the talker, CV syllable, and repetition number, respectively. For example, $O_{(1:51,1:34)}^{M1,ba,1}$ represents data for the first repetition of syllable /ba/ for talker M1. Each matrix has 34 columns that represent 34 frames and 51 rows that represent the optical channels (17 reflectors in a 3-D space). The physical Euclidean distance between a pair of consonants (C_1, C_2) with vowel context V for talker T was measured as follows:

$$PO_m^{T,C_1-C_2,V} = \sqrt{\sum_{j=1}^{2}\left(\sum_{n=1}^{34}\left(O_{m,n}^{T,C_1V,j} - O_{m,n}^{T,C_2V,j}\right)^2\right)}, \quad (2)$$

where m is the channel number (1-51), n is the frame number, and j is the repetition number. If all the Euclidean distances between the 23 consonants in a vowel context for talker T were put together, a 51 by 253 matrix can be obtained as $PO^{T,V}$ where each row represents a different optical channel and each column represents a different consonant pair.

We can ignore either talker effect or vowel context, and then two types of physical distances can be computed as:

$$PO_{m,n}^{ALL,V} = \sqrt{\left(PO_{m,n}^{M1,V}\right)^2 + \left(PO_{m,n}^{F1,V}\right)^2 + \left(PO_{m,n}^{M2,V}\right)^2 + \left(PO_{m,n}^{F2,V}\right)^2}, \quad (3)$$

$$PO_{m,n}^{T,aiu} = \sqrt{\left(PO_{m,n}^{T,a}\right)^2 + \left(PO_{m,n}^{T,i}\right)^2 + \left(PO_{m,n}^{T,u}\right)^2}, \quad (4)$$

where T can be talker M1, F1, M2, F2, or all talkers. For these distances, three subsets can be derived according to the reflector locations, which are physical distance matrices computed for the lip, cheek (chk), and chin (chn) reflectors, respectively.

2.3 Perceptual experiments

2.3.1. Viewers

The participants were six individuals with normal hearing, normal or corrected vision, English as a native language, and screened for average or better lipreading ability.

2.3.2. Procedure

Visual perception was assessed using high-quality video recordings as mentioned in Section 2.1.2. The experiments were conducted in a sound booth with a 19'' high-resolution color monitor for video presentations and a PC for controlling and recording viewers' responses. The videotapes (without sound) were presented to each viewer 10 times. Therefore, for each CV syllable, there were 480 responses (2 repetitions, 10 trials, 4 talkers, and 6 viewers). A detailed description of the experimental procedure can be found in [4].

2.3.3. Visual perception confusion matrices (VPCM)

Perceptual data consisted of six viewers' identifications of 23 consonants through lipreading each of the four talkers. Results were organized in different ways. The results were pooled for each talker and vowel context so that there are 12 confusion matrices (23x23). $V_{T,V}$ represents one confusion matrix where T is talker and V is the vowel context. In each of these confusion matrices, each stimulus has 120 responses (two repetitions, six viewers, ten presentations).

Again, we can ignore either talker effect or vowel context, and then two types of confusion matrices can be computed as:

$$V_{ALL,v}=V_{M1,v}+V_{F1,v}+V_{M2,v}+V_{F2,v}, \quad (5)$$

$$V_{T,aiu}=V_{T,a}+V_{T,i}+V_{T,u}, \quad (6)$$

A phi-square transformation was applied to these confusion matrices. The phi-square coefficient for an individual consonant pair ($C_1$and C_2) is independent of other consonants. Further, the phi-square coefficient has an advantage when there are response biases and asymmetries [8]. The resulting matrices represented the dissimilarity structure in visual perceptual space. These matrices were symmetric (the number of distances is 253 for the 23 consonants). We use a notation of $VD_{T,V}$ to represent visual consonant distances for talker T in a context V, which is a vector of 1 x 253. Again, we can derive 20 such vectors for the corresponding 20 raw confusion matrices.

2.3.4. Hierarchical clustering analysis (HCA) and phonemic equivalence classes

In this study, HCA [11, 12] generates an inverted tree structure in which phonemes join classes based on dissimilarity. At the lowest level of the structure, no phonemes are joined together. At each succeeding level, the most similar pair of classes is joined together. This continues until, at the highest level, all phonemes join a single equivalence class. An average-linkage-between-groups method was used to compute distances at each level in the hierarchy; two classes were joined if they had the minimum average between-class distance at that level.

Phonemic equivalence classes were chosen by finding the first level in which at least 75% of all the responses were within-class, similar to [6]. As an example, if the phonemes /b/ and /m/ were in the same class, then a /b/ response to a /m/ stimulus would be considered to be a within-class response.

3. RESULTS

3.1. Overall visual perception results

TABLE I. Lipreading accuracy for 23 consonants across talkers and vowel context.

Talker	Vowel context			
	C/a/	C/i/	C/u/	C/a, i, u/
M1	0.36	0.34	0.30	0.33
M2	0.39	0.34	0.32	0.35
F1	0.32	0.32	0.31	0.32
F2	0.37	0.37	0.32	0.35
All talkers	0.36	0.34	0.31	0.34

The results in Table I are somewhat similar to those reported in [7] (40% for /aCa/, 33% for /iCi/, and 24% for /uCu/); but lower than those reported in [8] (48% for CVs). In general, lipreading accuracy for C/a/ syllables is better than for C/i/ and C/u/ syllables. The lipreading accuracy for M2 and F2 (high sentence intelligibility talkers) is slightly higher than that of M1 and F2 (low sentence intelligibility talkers).

For visual speech perception, groups of phonemes that are indistinguishable visually are called 'visemes' [9]. Traditionally, consonants have been grouped into 12 categories (visemes) [2]: {h}, {k, g}, {y}, {□ □ t□ d□}, {r}, {l}, {s, z}, {t, d, n}, {□ □}, {f, v}, {w}, {p, b, m}. These visemes suggest that, in normal situations, it is difficult to distinguish phonemes within one viseme group. Lipreading accuracy was recomputed based on the 12 traditional visemes and is shown in Table II. Obviously, a higher accuracy would be obtained: 0.69 for C/a/ syllables, 0.66 for C/i/ syllables, and 0.58 for C/u/ syllables. Note that talker F2's speech had the highest lipreading accuracy based on the 12 viseme categories, although her visual sentence intelligibility was not the highest. This could be because the perceived visual categories for talker F2 tend to fall into the traditionally defined 12 viseme categories, while they did not for the other talkers.

TABLE II. Lipreading accuracy for 12 traditional viseme groups across talkers and the vowel context.

Talker	Vowel context			
	C/a/	C/i/	C/u/	C/a, i, u/
M1	0.64	0.64	0.52	0.60
M2	0.72	0.66	0.55	0.64
F1	0.65	0.61	0.60	0.62
F2	0.74	0.71	0.64	0.70
All talkers	0.69	0.66	0.58	0.64

3.2. Phonemic equivalence classes

TABLE III. Phonemic equivalence classes across talkers and vowel context.

VPCM	Phonemic equivalence classes
$V_{M1,a}$	(w) (m p b) (r f v) (h) (y l n k g θ ð) (t d s z ʃ ʒ tʃ dʒ)
$V_{M1,i}$	(w) (m p b) (r f v) (y k g h) (l n t d θ ð) (s z ʃ ʒ tʃ dʒ)
$V_{M1,u}$	(y w k g h) (m p b) (r f v) (l n t d θ ð) (s z ʃ ʒ tʃ dʒ)
$V_{M1,ahi}$	(w) (m p b) (r f v) (y l n t k d g h θ ð s z ʃ ʒ tʃ dʒ)
$V_{M2,a}$	(w r) (m p b) (f v) (θ ð) (y l n k g h) (t d s z ʃ ʒ tʃ dʒ)
$V_{M2,i}$	(w r) (m p b) (f v) (θ ð) (y l n t k d g h) (s z ʃ ʒ tʃ dʒ)
$V_{M2,u}$	(w r) (m p b) (f v) (θ ð) (y l n k g h) (t d s z ʃ ʒ tʃ dʒ)
$V_{M2,ahi}$	(w r) (m p b) (f v) (θ ð) (y l n k g h) (t d s z ʃ ʒ tʃ dʒ)
$V_{F1,a}$	(w r) (m p b) (f v) (θ ð) (y l n t k d g h s z ʃ ʒ tʃ dʒ)
$V_{F1,i}$	(w r) (m p b) (f v) (θ ð) (y l n t k d g h s z ʃ ʒ tʃ dʒ)
$V_{F1,u}$	(w r m p b) (f v) (θ ð) (y l n t k d g h s z ʃ ʒ tʃ dʒ)
$V_{F1,ahi}$	(w r) (m p b) (f v) (θ ð) (y l n t k d g h s z ʃ ʒ tʃ dʒ)
$V_{F2,a}$	(w) (m p b) (r f v) (y l n t k d g h θ ð s z ʃ ʒ tʃ dʒ)
$V_{F2,i}$	(w) (m p b) (r f v) (θ ð) (y l n k g h) (t d s z) (ʃ ʒ tʃ dʒ)
$V_{F2,u}$	(m p b) (r f v) (θ ð) (y w l n t k d g h s z ʃ ʒ tʃ dʒ)
$V_{F2,ahi}$	(w) (m p b) (r f v) (θ ð) (y l n k g h) (t d s z) (ʃ ʒ tʃ dʒ)
$V_{ALL,a}$	(w) (m p b) (r f v) (θ ð) (y l n k g h) (t d s z ʃ ʒ tʃ dʒ)
$V_{ALL,i}$	(w r) (m p b) (f v) (θ ð) (y l n t k d g h) (s z ʃ ʒ tʃ dʒ)
$V_{ALL,u}$	(w) (m p b) (r f v) (θ ð) (y l n k g h) (t d s z ʃ ʒ tʃ dʒ)
$V_{ALL,ahi}$	(w) (m p b) (r f v) (θ ð) (y l n k g h) (t d s z ʃ ʒ tʃ dʒ)

Table III lists the phonemic equivalence classes for each lipreading condition (different talkers and vowel context). The table clearly shows that the number of visemes viewers can identify is much smaller than 12. For example, talker M2, with the highest sentence intelligibility rating, had six visemes. Talker F1, with the lowest sentence intelligibility rating, had five visemes for C/a/ and C/i/ syllables and four viseme groups for C/u/ syllables. Talker F2, with a medium-high sentence intelligibility rating, had only four viseme groups for C/a/ syllables and four visemes for C/u/ syllables, but had the highest number of visemes for C/i/ syllables, seven. However, lipreading accuracy for F2's C/i/ syllables was not higher than for C/a/ syllables (Table I). It appears that visual sentence intelligibility does not solely depend on the number of visemes. It may also depend on which consonants are perceived correctly. For example, the frequencies of /w/, /□, and /□ are usually high in

words so that it is important to identify these consonants correctly [10]. As shown in Tables I and II, C/u/ syllables were more difficult to lipread than C/a/ and C/i/ syllables, and this was reflected in Table III because in general fewer phonemic equivalence classes were obtained for the confusion matrices $V_{T,u}$ than for $V_{T,a}$ and $V_{T,i}$.

3.3. Predicting visual perception from physical measures

Multiple linear regression techniques were used to assess the relationship between visual consonant perception and physical measures so that the factors contributing to visual sentence intelligibility could be examined. The physical measures for consonants, and the physical distances between consonants, employed here were discussed in Section 2.2.2. For example, in the vowel /a/ context, these measures are referred to as $PO^{T,a}$ (51x253, 17 reflectors), $PO^{T,a}_{lip}$ (24x253, 8 reflectors on the lips), $PO^{T,a}_{chk}$ (18x253, 6 reflectors on the cheeks), and $PO^{T,a}_{chn}$ (9x253, 3 reflectors on the chin). In Section 2.3.3, we described how to obtain visual distances ($VD_{T,v}$).

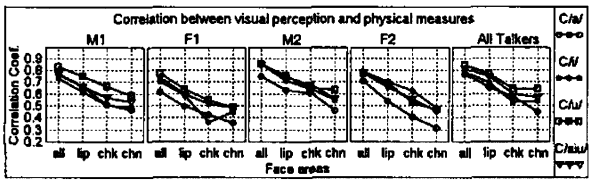

Figure 4. Predicting visual perception from physical measures

Fig. 4 shows the results for the correlation analysis between perceptual and physical measures. For all talkers, the correlations were 0.77 for C/a/, 0.78 for C/i/, and 0.84 for C/u/, and 0.81. Across all contexts and reflectors, the correlations were 0.77 for M1, 0.74 for F1, 0.85 for M2, and 0.78 for F2. Talker M2, with the highest sentence intelligibility, had the highest correlations, and Talker F1, with the lowest sentence intelligibility, had the lowest correlations. The two males tend to have higher correlations than the two females, which could be due to the two males having larger faces and larger movements than the two females. In general, about 63% of the variance in visual consonant confusion was accounted for by physical measures. As for vowel context, C/u/ syllables yielded higher correlations between perceptual and physical measures than C/a/ and C/i/ syllables even though the movements for /u/ are smaller. This could be due to the fact that the movements for /u/ are more concentrated around the lips. Fig. 4 also shows that both the lips and the cheeks were important for visual perception. Considering the panel with data for all talkers, if one uses lip data only, the correlation across all three vowels C/aiu/ is 0.74. Including the cheek and chin data improves the correlation to 0.81. In addition, cheek data yielded higher correlations than the chin data suggesting that various cheek movements help identify visual gestures that are important for visual perception.

4. SUMMARY AND CONCLUSIONS

In this paper, we examined the relationship between perceptual and optical measures for visual consonants. There are two main findings in the study. First, visual perception experiments revealed that phonemic equivalence classes were fewer than 12, talker and vowel context dependent, and related to talkers' sentence intelligibility. In general, higher sentence intelligibility corresponded to more phonemic equivalence classes.

The study also showed high overall correlations (0.77, 0.74, 0.85, and 0.78 for M1, F1, M2, and F2, respectively) between perceptual and physical measures. Combining the three physical measures (lips, cheeks, and chin) resulted in higher correlation than using only the lips. Of the face movements, the lips and cheeks are more informative for visual perception, than the chin. As for vowel context, the /u/ context yields the highest correlations. The correlations between perceptual and physical measures were also a function of visual sentence intelligibility ratings and talker. High visual sentence intelligibility resulted in higher correlations and so did male speech. During lipreading, viewers may have tried to guess from time to time. If a talker's sentence intelligibility is low, then lipreading may be vulnerable to noise (guessing) and this would result in lower correlations.

5. ACKNOWLEDGEMENTS

This research was supported in part by the NSF KDI award 9996088. We wish to acknowledge the help of B. Chaney, P. Barjam, J. Yarbrough, S. Mattys, and T. Cho.

6. REFERENCES

[1] A. Montgomery and P. Jackson, "Physical characteristics of the lips underlying vowel lipreading performance,", JASA, vol. 76, pp. 2134-2144, 1983.

[2] P. Kricos and S. Lesner, "Differences in visual intelligibility across talkers," The Volta Review, vol. 84, pp. 219-225, 1982.

[3] C. Benoît, T. Guiard-Marigny, B. Le Goff, and A. Adjoudani, "Which components of the face do humans and machines best speechread?" In Stork and Hennecke (Eds.), *Speechreading by Humans and Machines*, pp. 315-328, 1996

[4] J. Jiang, A. Alwan, E. Auer, and L. Bernstein, "Predicting visual consonant perception from physical measures," EUROSPEECH'01, vol. 1, pp. 179-182, 2001

[5] L. Bernstein, E. Auer, B. Chaney, A. Alwan, and P. Keating, "Development of a facility for simultaneous recordings of acoustic, physical (3-D motion and video), and physiological speech data," JASA, vol. 107, p2887, 2000.

[6] B. Walden, R. Prosek, and A. Montgomery, "Effects of training on the visual recognition of consonants," JSHR, vol. 20, pp.130-145, 1977

[7] E. Owens and B. Blazek, "Visemes observed by hearing-impaired and normal hearing adult viewers," JSHR, vol. 28, pp. 381-393, 1985

[8] P. Iverson, L. Bernstein, and E. Auer, "Modeling the interaction of phonemic intelligibility and lexical structure in the audiovisual word recognition," Speech Comm., vol. 26, pp. 45-63, 1988

[9] C. Fisher, "Confusion among visually perceived consonants," JSHR, vol. 11, pp. 796-804, 1968

[10]E. Auer and L. Bernstein, "Speechreading and the structure of the lexicon: computationally modeling the effects of reduced phonetic distinctiveness on lexical uniqueness," JASA, vol. 102(6), pp. 3704-3710, 1997

[11]M. Aldengerfer and R. Blashfield, *Cluster analysis*. Beverly Hills and London: Sage Pubns, 1984

[12]SPSS® Base 9.0 User's Guide.